Machine Learning and Political Events

A Novel Data Set on Latin American Cabinets

Bastián González-Bustamante [1] [2]

[1]Department of Politics and International Relations [2]St Hilda's College

Data Mining and Machine Learning

Based on Natural Language Processing (NLP) and the text-as-data approach, we present an entirely novel data set constructed by applying data mining and machine learning techniques.

The data set contains detailed information on cabinet turnover in 12 Latin American democracies from the mid-1970s to date.

Calls for the Resignation of Ministers

We used press archives that were digitised with data mining algorithms using optical character recognition (OCR). Then, using machine learning models, we were able to identify ministerial resignation calls during the period.

Although the literature has identified that protests, economic crises, scandals of different sorts, among other stochastic events, do affect stability (Martínez-Gallardo, 2014), assessing all these possible random shocks is complex and empirically costly (Berlinski et al., 2010). Therefore, cues such as calls for the resignation of ministers may be empirically efficient indicators.

This indicator is entirely unprecedented in Latin American presidential systems and offers a number of potential applications beyond the study of coalitions and ministerial turnover, opening new avenues for future research on public opinion, political polarisation or public policy.

Optical Recognition on Press Archives

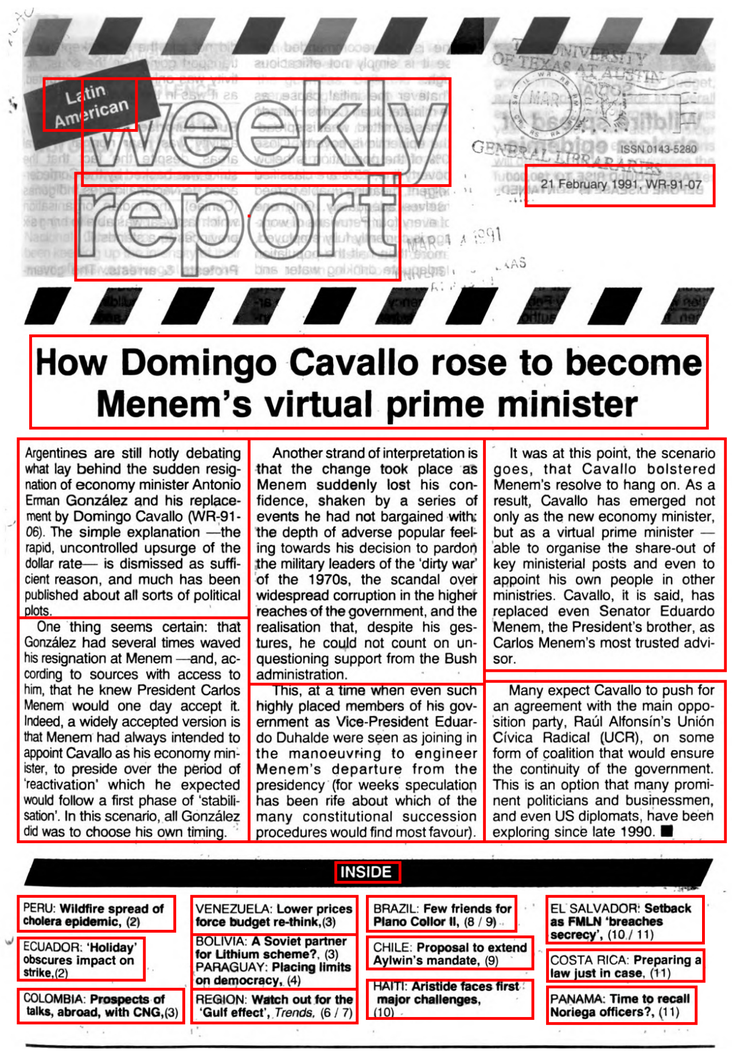

Figure 1. Tesseract's Layout Analysis Module

Machine Learning Models

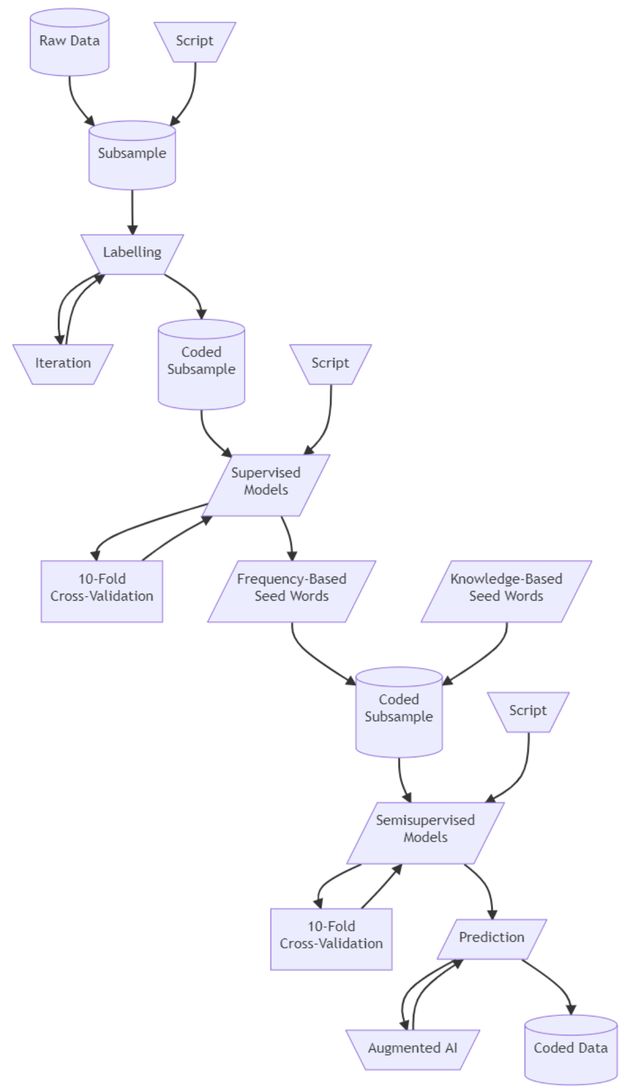

Figure 2. Models Training Pipeline

Trained Algorithms

- Naive Bayes (NB)
- Kernel Linear Support Vector Machine (SVM)
- Kernel Gaussian Radial Basis Function (RBF)
- Random Forest with 100 trees (RF100)
- Random Forest with 500 trees (RF500)
- Extreme Gradient Boosting (XGBoost)

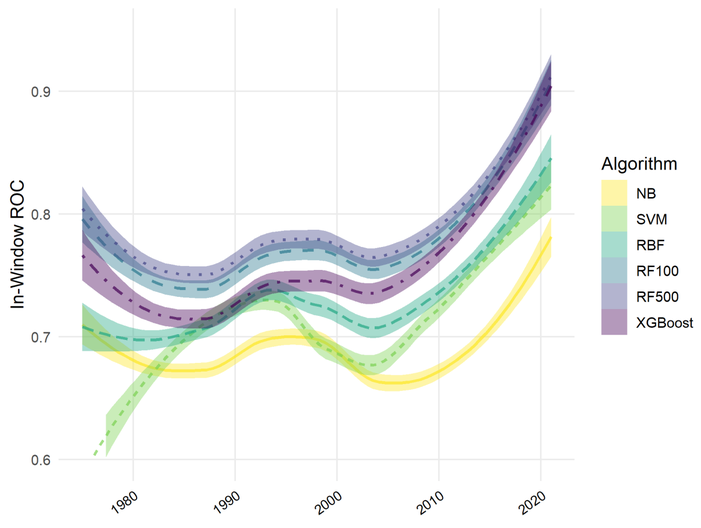

Figure 3. Goodness-of-Prediction. *Note*: ROC indicators were 10-fold cross-validated in a five-year rolling window training set between 1975 and 2021.

Plausibility and Validity Checks

We present two practical applications with nonparametric Kaplan-Meier survival estimations and topic modelling to identify dominant issues in public opinion in the region between 1975 and 2021.

Survival Estimations

We can use $F(t) = 1 - S(t) = Pr(T \leq t)$ to obtain the incidence of ministerial exits during T by considering $S(t)$ as a survival function. In this context, following Austin et al. (2016), Kaplan-Meier nonparametric estimations were fitted based on the number of ministers leaving the cabinet (or were reallocated) over the number at the beginning of the presidential term.

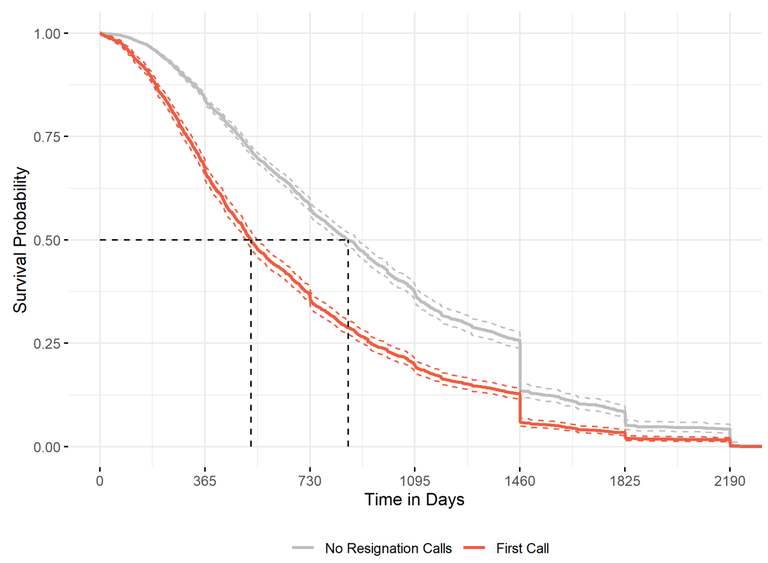

Figure 4. Survival for Ministerial Turnover

Topic Modelling

We have used Keyword Assisted Topic Modelling (keyATM; see Eshima et al., 2022). This involves specifying certain keywords to generate labels prior to topic modelling.

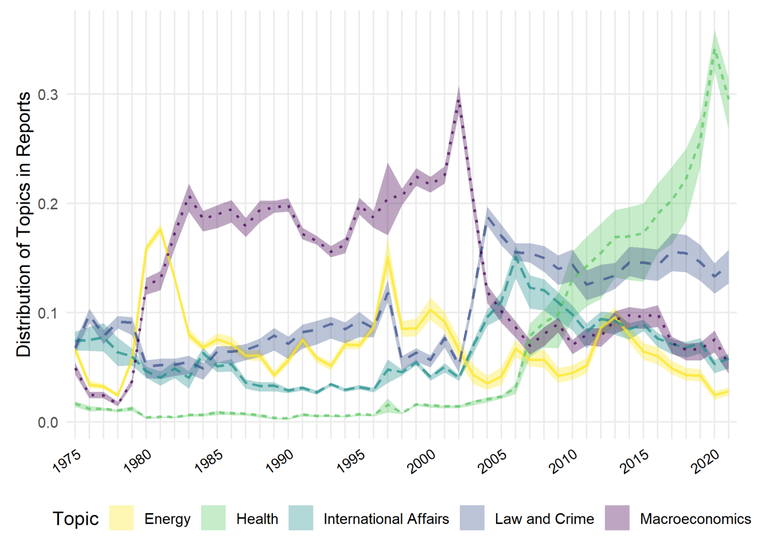

Figure 5. Agenda Topics in Latin America (1975-2021)

References

Austin, P. C., Lee, D. S., and Fine, J. P. (2016). Introduction to the Analysis of Survival Data in the Presence of Competing Risks. *Circulation*, 133(6):601–609.

Berlinski, S., Dewan, T., and Dowding, K. (2010). The Impact of Individual and Collective Performance on Ministerial Tenure. *The Journal of Politics*, 72(2):559–571.

Eshima, S., Sasaki, T., and Imai, K. (2022). Keyword Assisted Topic Models. Package, Harvard University and Massachusetts Institute of Technology. Package 'keyATM' version 0.4.1.

Martínez-Gallardo, C. (2014). Designing Cabinet: Presidential Politics and Ministerial Instability. *Journal of Politics in Latin America*, 6(2):3–38.

Download this poster from GitHub